**Curation and Expansion of Human Phenotype Ontology for Defined Groups of Inborn Errors of Immunity**

# Methods

# *Organization of working groups, working group participants*

# Working groups were established following the 2017 IUIS classification categories (*27*). Based on the participants expertise, the working groups addressed the following IUIS classification categories: diseases affecting cellular and humoral immunity (IUIS Table 1), predominantly antibody deficiencies (IUIS Table 3), diseases of immune dysregulation (IUIS Table 4), and autoinflammatory disorders (IUIS Table 7). Headed by a group lead, themselves an expert in the field of the specific diseases, each working group was accompanied by an additional member (coordinator) from the Ludwig Boltzmann Institute for Rare and Undiagnosed Diseases (LBI-RUD) to facilitate coordinated communication and organization. Working groups had between 5 and 9 members with disease-group-specific standard operational procedures in place to facilitate the workflow. Coordinators initiated and communicated general agendas within each working group, implemented disease-group-specific tasks, hosted remote meetings and communicated results and subsequent actions via email.

# Table S1: Organization and working group participants

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Organization & coordination** | | | **Specific working groups** | | | | |
| **Initiative**  **leaders** | **Main  coordinators** | **Group**  **coordinators** | **Diseases affecting innate and adaptive immunity**  **(IUIS Table I)** | **Antibody deficiencies**  **(IUIS Table III)** | **Diseases of immune dysregulation**  **(IUIS Table IV)** | **Autoinflammatory diseases**  **(IUIS Table VII)** |
| Kaan Boztuga,b,c,ak,al | Matthias Haimela,b,c | Raúl Jiménez Herediaa,b,c | Sophie Hambletonaf | Esther   de Vriesag,ah | Catharina Schuetzac | Anna Simonad |
| Marielle van Gijnaj | Julia Pazmandia,b,c | Jasmin Dmytrusa,b,c | Kimberly C. Gilmourm | Siobhan Burnsx | Shahrzad Bakhtiaro | Paul Broganl,m |
| Peter Robinsonai |  | Sevgi Köstel Bala,b,c | Fabian Hauckaa,ab | Johannes Trückr | Beata Derfalvip | Marco Gattornok |
| Christoph Bocka,c,ae |  | Samaneh Zoghia,b,c | Francesco Saettiniq | Jacques Rivières,t | Markus G. Seidelw | Paul van Daeled |
|  |  | Julia Pazmandia,b,c | Mirjam van der Burgu,v |  |  | Tracy A Briggse,f |
|  |  |  | Maaike A. A. Kustersl,m |  |  | Carine Woutersg,h |
|  |  |  | Reem Elfekyl,m |  |  | Brigitte Bader-Meunieri,j |
|  |  |  | Elisabeth Salzera,b,c,al |  |  | Florence Aeschlimani,j |
|  |  |  |  |  |  | Roberta Caorsik |
|  |  |  |  |  |  | Despina Eleftherioul,m |
|  |  |  |  |  |  | Esther Hoppenreijsn |

#### Expansion and restructuring of disease related branches of the HPO tree

OMIM (*5*) (if unavailable, OrphaNet (*6*)) identifiers were used as starting points to refer to Inborn Errors of Immunity (IEI)s. For each disease, available Human Phenotype Ontology (HPO) terms and the ontology structures were extracted from the v2018-06-13 HPO disease annotation and ontology release (https://hpo.jax.org), and Excel documents were prepared summarizing the annotations per disease. Each document contained the currently available HPO terms associated with one disease, upstream terms of these current HPO terms organized in a tree structure. Both the correctness of available terms and the ontology structure associated with the terms was assessed. Disease-specific HPO restructuring was discussed within the working groups and the results debated among all participants. The suggested changes were summarized electronically in Excel documents or pictures of flipchart drawings by the main coordinators before being submitted to HPO. Additionally, missing terms describing pulmonary and gastro-intestinal complications of Primary Antibody Deficiency (PAD) were discussed during teleconferences and thereafter submitted to update the HPO ontology. All results of the restructuring are detailed in Supplementary Document S1&S2.

#### Standardized re-annotation of rare, genetically diagnosed diseases

A standardized, semi-automated reannotation process was developed in order to consistency annotate all IEIs (over 300 different diseases in OMIM) with HPO terms across working groups.

In the first step, working groups collected a minimum of two articles in portable document format (PDF) that adequately illustrated the phenotypic spectrum of each disease. In the second step, the text was extracted from the PDF files using the content analysis tool Apache Tika (https://tika.apache.org/) from the python package tika (version 1.19). Text sections associated with HPO terms were identified by applying an ontology-guided machine learning tool, the Neural Concept Recognizer (NCR), with default settings as previously published (*17*), trained on the v2019-06-03 HPO ontology release. The NCR was selected due to the ability to work with HPO, utilize the hierarchy information and semantic similarity for improved identification of HPO terms, and the robust performance on a published PubMed article abstracts dataset with manual HPO annotations (*17,R1*). The identified HPO terms and term frequencies were collected for each article and further summarized per disease. These disease specific summaries were prepared as Excel documents, where HPO terms were ranked by frequency across articles. Highlighted HPO terms indicated already available annotations in the v2019-06-03 HPO disease annotation release on the same HPO branch (of this specific or more/less specific HPO term). These Excel documents were distributed to the working groups and evaluated with two-tier expert review. For the two-tier expert review, each disease was reviewed by at least two independent experts. The experts were asked to indicate if an HPO term was either a true phenotype present in the disease, or a false positive association. In addition to evaluating the HPO terms identified in the articles, the experts evaluated existing HPO terms from the HPO annotation release as well. Further phenotypes not identified by the previous two steps (identified in articles or available in the HPO annotation release) were suggested by experts as additional terms (HPO terms if available or free text) to cover the full phenotypic spectrum of the diseases.

In case of any disagreement in the evaluation, a consensus discussion between the two experts for that particular disease was scheduled. If after the second-tier overview between the two experts there was still no agreement, these were discussed by the whole group for overall consensus, defined as at least 80% agreement amongst the expert group. The consensus of the expert evaluations was collected in standardized Excel documents. These consensus Excel documents per disease were integrated by the main coordinators at LBI-RUD with all diseases across working groups. The full list of reannotated diseases available in Supplementary Document S3. The list of reannotated terms for each disease is available in Supplementary Document S4.

#### Standardized re-annotation of genetically undiagnosed diseases

Literature describing the phenotypic characteristics of the various PAD subtypes without a known monogenetic defect were collected and reviewed by the PAD-subgroup members. The ontology-guided machine learning tool was run as described above. Each HPO identifier (either identified in an article or available in the HPO annotation release) was annotated with: true/false/exclusion criteria. In case of a true phenotype, the observed frequency within patients was assessed and noted down as well. The frequencies correspond to the following representation in patients: Common = Frequent (79-30%); sometimes = Occasional (29-5%); rare = Very rare (<4-1%).

#### Preparation of disease annotation data for similarity measures

The list of existing annotations per disease (later referred as “HPO-disease-annotations”) was obtained by extracting the HPO terms available per disease from the v2020-03-27 HPO release. Redundant terms (as defined as terms, where more specific terms are already linked and available to the disease) were removed from each disease annotation by applying the *minimal\_set* function, based on the v2020-03-27 HPO release ontology structure. To obtain the reannotated set of annotations (detailed above, later referred to as “reannotated-disease-annotations”), the list of reviewed and evaluated disease annotations was extracted from the working groups. Redundant terms were removed with by applying the *minimal\_set* function.

#### Disease-disease similarity measures

Similarity measures of diseases based on both disease annotations sets - “HPO-disease-annotations” and “reannotated-disease-annotations” - was carried out by the R package ontologyX (*18*), with default settings applying the Lin similarity measure. A phenotypic similarity matrix of disease similarity data was calculated for both sets of annotations using the *get\_sim\_grid* method with default parameters. Diseases were clustered based on this similarity matrix using euclidean distances. Hierarchical clustering was performed and visualized with *ggtree* using the *R* packages ggtree v1.14.6 (*19*) and ape v5.2 (*20*).

*Patient-disease similarity measures*

For each patient, HPO terms from clinical synopses were extracted with the Neural Concept Recognizer (*17*), then reviewed and expanded by an expert clinician. The semantic similarity of the extracted HPO terms per patient was compared to all diseases in both disease annotation sets, “HPO-disease-annotations” and “reannotated-disease-annotations” (see above), using the R package ontologyX (*18*), with default settings using the Lin similarity measure, by applying *get\_profile\_sim*. The statistical significance of the difference between cohort-wide similarity scores to genetic diagnosis using “HPO-disease-annotations” (therefore before reannotation) and “reannotated-disease-annotations” (therefore after reannotation) was assessed by a Student T-test using the R package ggpubr v0.2.999. Next, per patient, all diseases were ranked according to their similarity scores to the list of patient HPO terms. Diseases with the highest similarity received the lowest rank. The statistical significance of the difference of rank comparing the two annotations - “HPO-disease-annotations” (therefore before reannotation) and “reannotated-disease-annotations” (therefore after reannotation) - was assessed by Student T-test using the R package ggpubr v 0.2.999. A detailed list of patients and their similarity to their genetic diagnosis can be found in Supplementary Document S4.

#### List of Human Phenotype Ontology Resources:

|  |  |
| --- | --- |
| **Data** | **Accessible from** |
| Current version of the HPO ontology | https://hpo.jax.org/app/download/ontology |
| Current HPO annotation of diseases | https://hpo.jax.org/app/download/annotation |
| List of HPO ontology versions | https://github.com/obophenotype/human-phenotype-ontology/releases https://bioportal.bioontology.org/ontologies/HP |
| v2018-06-13 HPO ontology release | http://purl.obolibrary.org/obo/hp/releases/2018-06-13/hp.owl |
| v2019-06-03 HPO ontology release | http://purl.obolibrary.org/obo/hp/releases/2019-06-03/hp.owl |
| v2020-03-27 HPO ontology release | http://purl.obolibrary.org/obo/hp/releases/2020-03-27/hp.owl |
| HPO disease annotation archive file | https://archive.monarchinitiative.org/hpo-archive/20210126\_jenkins\_jobs.tar.gz |
| v2018-06-13 HPO annotation release | jobs/hpo.annotations/builds/1254/archive/misc (in HPO annotation archive file) |
| v2019-06-03 HPO annotation release | jobs/hpo.annotations/builds/1266/archive/misc (in HPO annotation archive file) |
| v2020-03-27 HPO annotation release | jobs/hpo.annotations/builds/1271/archive/misc (in HPO annotation archive file) |
| Current HPO disease annotations | https://ci.monarchinitiative.org/view/hpo/job/hpo.annotations/ |

#### List of Supplementary Documents:

Supplementary Methods

Supplementary Document S1: HPO tree restructuring and the list of new terms

Supplementary Document S2: Summary of diseases reannotated

Supplementary Document S3: List of all terms per disease after reannotation

Supplementary Document S4: List of cases used for phenotype to diagnosis matching

**References**

1. Groza T, Köhler S, Doelken S, Collier N, Oellrich A, Smedley D, et al. Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database (Oxford)*. 2015;2015.